Study of Mitochondrial RNA Mutations and their role in Alzheimer's Disease

Macrina M. Lobo, Master of Science Candidate
Columbia University, New York City, New York, United States

Abstract

Mutations often cause or are caused by disease. Understanding this relation may provide insights into disease pathogenesis. Next generation methods have made gene sequencing data easily available. Genome wide association studies capitalize on this huge amount of data to obtain information such as whether a particular mutation is more prevalent in a diseased population. Research has linked mitochondrial variants to Alzheimer's disease (AD). I attempt to gain insights into this linkage using raw mitochondrial RNA sequencing data obtained from 146 healthy and diseased brain cells. I test several hypotheses regarding the factors influencing AD. I use knowledge gained from hypotheses testing to build a logistic regression model which uses age, gender and extent of mutation in a cell to detect the presence of AD.

1. Introduction

Several age-related diseases such as Alzheimer's disease (AD) result from damage to mitochondria [5]. Recent work has studied the relation between SNPs and AD [4]. However, to the best of my knowledge, few GWAS studies have focused on mitochondria variants. [6] focused on the UCP gene family. [3] studies mitochondrial variants associated with lipid profiles and late onset diseases.

Our contributions are manifold: (1) As described later, mitochondrial (mt) SNPs must be studied differently from nuclear SNPs due to heteroplasmy. This is one limitation of [4,6]. I hope to use allele counts at each position in a manner similar to [3] to solve this. (2) Quantitative trait analysis at the cell level using statistical measures. I wish to answer questions such as (a) Are age/gender related to AD? (b) Are the total number of mutations in a cell related to AD? (c) What about allelic counts and AD? (3) Study rare and common mutations and their relationship with AD. (4) determine if certain genotypes influence the number of mitochondria in a cell.

2. Methods

2.1 Dataset

AD is a brain related disease hence mitochondrial sequencing data from brain cells is required. I found it difficult to obtain brain mtDNA data and used mtRNA data from Religious Orders Study and Memory and Aging Project (ROS/MAP) Study. The data was obtained from ROSMAP by the Taub Institute for research on Alzheimer's and the Aging Brain, CUMC and is private. A set comprising RNA sequencing data from 146 healthy and diseased brain cells of different human subjects is used here. Metadata includes age, gender, post-mortem interval (pmi).

2.1.1 Dataset Challenges

Each mitochondria of an individual contains 2-10 mtDNA copies [1] and different mitochondria may have different genotypes [2]. Different cells also have different number of mitochondria. This heterogeneity poses a challenge to mtRNA sequencing data analysis. Sequencing and mapping tools provide diploid counts though the mt-genome is haploid. This is because the apparent 0/1 genotype in reality represents the presence of both the dominant and recessive allele at that loci within the cell. Thus, a single cell can exhibit the two alleles in different percentages at a single locus. The tertiary genotype for a given variant "homozygote reference/reference" (0/0), "heterozygote reference/mutant" (0/1) and "homozygote mutant/mutant" (1/1) does not hold in this case and one should consider a more continuous, quantitative characterization.

2.2 Processing of raw data

I obtained clean data following the procedure in **Figure 1**. I used SRATOOLS (namely fastq-dump) to perform paired end sequencing. For bowtie2 [7], I used the local alignment option with higher sensitivity at the cost of longer run time. The local option maximizes the alignment even at the cost of ignoring reads at the ends. For samtools [8], I ignored insertions and deletions (indels) and extracted only mitochondrial chromosome (16569 positions) information. Samtools has a Base Alignment Quality (BAQ) index. Its aim is to rule out false SNPs caused by nearby indels but if used, it misses several true positives as well. Hence I disabled it. I wrote a shell script to carry out the steps on all the 146 files.

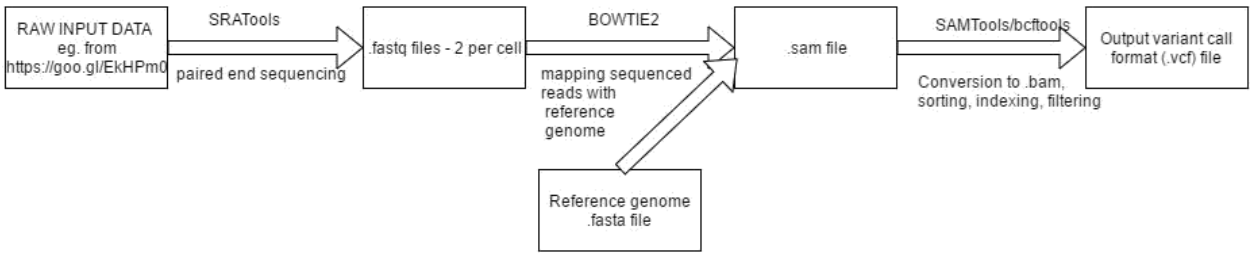


Figure 1. Raw data processing pipeline

I removed cells with missing AD or incomplete sequencing information to obtain information from 39 AD and 71 healthy cells. The output file is human readable. The .vcf file contains information such as chromosome name/number (in our case mitochondria), position of the base on the chromosome, reference and alternate sequence, quality of the mapping, raw read depth, scaled genotype likelihoods, observed genotype (0/0, 0/1 or 1/1) and number of reads matching reference and alternate allele on the + and – strands. This last term is less than raw read depth since it is obtained after applying the default samtools filter for removing low quality read counts. The quality is specified in the 'quality' field. For this work, I used only the position, observed genotype and number of reads matching reference and alternate allele on the + and – strands. It might be interesting to consider scaled genotype likelihoods in the future.

To deal with the heteroplasmy problem, [3] proposed taking the log of allelic ratios. In order to obtain an allelic ratio count for each cell, I experimented with (1), (2) and (3)

$$\log_2 \frac{\sum_{positions}(number\ of\ reads\ matchinf\ reference\ allele\ on{+}and{-}strands)}{\sum_{positions}(number\ of\ reads\ matching\ alternate\ allele\ on{+}and{-}strands{+}\in)} \quad (1)$$

$$\frac{\sum_{positions}(number\ of\ reads\ matchinf\ reference\ allele\ on{+}and{-}strands)}{\sum_{positions}(number\ of\ reads\ matching\ alternate\ allele\ on{+}and{-}strands{+}\in)} \quad (2)$$

$$\sum_{position} \log_2 \frac{(number\ of\ reads\ matching\ reference\ allele\ on{+}and{-}strand{+}\in)}{(number\ of\ reads\ matching\ alternate\ allele\ on{+}and{-}strand{+}\in)} \quad (3)$$

where is a small constant (set to 0.001 here) to ensure the terms don't get to infinity. The summation over positions represents the 16569 positions in the mitochondrial chromosome obtained in this study.

2.3 Hypotheses testing

The aim of this work is to establish relationships between AD and different observed variables. I did this using hypothesis testing. A table listing the aim, null hypothesis and method used is shown below (Table 1).

	Aim: Test the relation between	Test	Null hypothesis: There is no difference between
1	Age and AD	2-sample t-test	Ages of healthy and AD subjects
2	Total number of mutations and AD	2-sample t-test	Number of mutations in healthy and AD subjects
3	allelic ratios and AD (equation (1), (2) and (3))	2-sample t-test	Number of mutations in healthy and AD subjects
4	Gender and AD	Fisher's exact test	Gender of AD and healthy subjects i.e. AD does not more commonly occur in subjects of a particular gender
5	Age and number of mutations	Pearson's correlation coefficient; no significance tested	-
6	Post mortem interval and AD	2-sample t-test	PMI in AD and healthy subjects
7	Rare mutation count and AD	2-sample t-test	Rare mutation count in AD and healthy subjects

Table 1. Hypothesis Tests

The t-tests referred to assume unequal sample size and equal variance though variances are in reality

unequal. This is because the unequal variance version requires a large number or perfectly Gaussian distributed samples to obtain correct results.

2.4 Logistic Regression

I identified the features affecting AD from the hypothesis tests and built a logistic regression model to predict the presence of AD from the features. I studied the p-values of the regression coefficients to assess their significance. I randomly divided the data into training and testing sets (approx. 70% training). I trained the model on the training set and computed accuracy on the testing set with a probability threshold for the logistic regression model output. I also used leave one out cross validation (LOOCV) to validate across a range of values for the threshold parameter. I experimented with probit regression as well but the results were almost identical to logistic regression.

2.5 Dealing with rare mutations

I computed the total number of mutations (0/1 + 1/1) across all cells at each position. Since the AD information is not required for this step, the number of useable cells was reduced from 146 to 127 instead of 110. I classified a mutation as rare if it occurred in less than 3 cells and thus obtained a table of rare and common mutations. Since very few rare mutations occur per cell (across all positions), I hypothesized that a count of total rare mutations in the cell would bear some relation with AD. Hence I performed Test 7 in Table 1.

2.6 Dealing with common mutations

For each common mutation, I hypothesized that it has an effect on AD. Choosing the null hypothesis that AD and healthy individuals had similar common mutation distributions, I performed a 2 sample t-test separately for each of the common mutations. My common mutation information is a binary indicator matrix with each element indicating whether the cell (row) had that particular common mutation (column).

2.7 Genotype and number of mitochondria

I propose to determine if certain genotypes/ mutations lead to an increase in the number of cell mitochondria. I do this by taking an estimated mitochondria count in each cell. I represented this by $estimated\ number\ of\ mitochondria\ in\ a\ cell = \frac{\sum_{positions} total\ number\ of\ reads\ at\ that\ position}{number\ of\ positions}$. Based on this, I compute simple statistics. I also compute the Spearmann's correlation coefficient between mutation information at each position and the estimated number of mitochondria. Here, too, mutation information is a binary indicator indicating the presence or absence of a mutation at a particular position. Rare as well as common mutations are considered. I also perform a t-test to study the difference between the number of mitochondria in AD as well healthy subjects.

3. Results

3.1 Processing data pipeline

I computed some simple statistics on the raw data obtained from the pipeline in **Figure 1**. They are listed in **Table 2** and **Table 3**.

	Age	# males	# females	0/1 counts	1/1 counts	Log allelic ratio (according to eqn (3))	Rare mutation count
Total	3440.47	15	24	159	115	6665770	64
Average	88.2	0.33	0.38	4.08	2.95	170917.2	1.64

Table2. Statistics for AD subjects

	Age	# males	# females	0/1 counts	1/1 counts	Log allelic ratio (according to eqn (3))	Rare mutation count
Total	6027.27	31	40	302	177	11998637	57
Average	84.89	0.67	0.62	4.25	2.49	168994.9	0.80

Table 3. Statistics for healthy subjects

In **Table2**, average number of males (or females) represents the ratio of the total number of male (or

female) subjects with AD to the total number of male (or female) subjects and similarly in **Table 3**.

As described earlier, the 0/1 and 1/1 genotype counts are obtained due to apparent diploid nature of the mt-genome. In reality the mitochondrial genome is haploid and using allelic ratios seems more appropriate.

3.2 Hypotheses testing

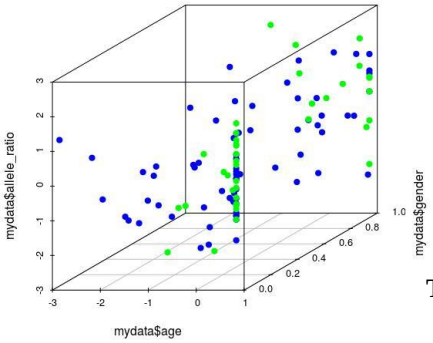
The results of the hypothesis tests for a 95% confidence interval are summarized in **Table 4**. The order of tests is the same as that in **Table 1**.

	Test	t	p-value	95% confidence interval
1	Age and AD	3.76	0.0003	[1.57, 5.08]
2	Total number of mutations 0/1 + 1/1 and AD	0.61	0.5442	[-0.63, 1.19]
3	Allelic ratios and AD (acc to eqn (1))	-0.99	0.3229	[-0.14, 0.05]
4	Allelic ratios and AD (acc to eqn (2))	1.39	0.1689	[-57177953, 322385803]
5	Allelic ratios and AD (acc to eqn (3))	2.40	0.0183	[332.22, 3512.35]
6	Gender and AD	-	0.6875	[0.33, 1.92]; odds ratio: 0.81
7	Age and number of mutations	-	-	Positive correlation: 0.1
8	Pmi and AD	-0.40	0.6922	[-2.18, 1.45]
9	Cell-level rare mutation count and AD	2.23	0.0278	[0.09, 1.58]

Table 4. Hypothesis Testing Results

3.3 Logistic regression results

A scatter plot of the data is shown in **Figure 2**. I used age, log allelic ratios according to equation (3) and gender as features to train a logistic regression model. At first, I used the entire dataset to train a logistic regression model. The estimates and p-values of the coeffcients are listed in T**able 5**.



	estimate	p-value
intercept	-0.8518	0.0075
age	0.9506	0.0022
gender	0.1701	0.7239
Log allele ratio	0.4400	0.0579

Table 5. Estimates and p-values of logistic regression model coefficients

Figure 2. Scatter plot. AD patients are in green while healthy subjects are in blue.

I then split the data into training and test set at approximately 70%-30% to obtain the accuracy of prediction. Setting the threshold probability for the logistic model at 0.5 and repeating the procedure 10 times (since the training and testing set were selected randomly) I obtained a highest accuracy of 81.82% and an average of accuracy of 72.73%

To study the behavior of the model as the threshold probability is varied, I performed cross validation. For small size datasets and limited tuning parameter values, leave-out-one-cross-validation (LOOCV) gives good validation performance and hence I used it here. I varied the probability in steps of 0.1 and obtained 11 values from 0 to 1. For each threshold parameter, using LOOCV, I obtained test results for each of the data points. I thus obtained a single true positive and false positive rate for each parameter value. I used these to plot the ROC curve shown in **Figure 3**.

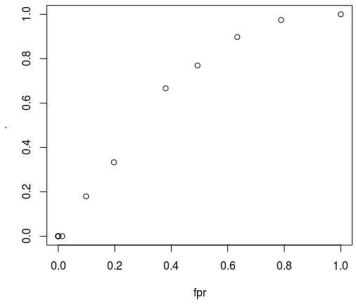


Figure 3. ROC curve is obtained for LOOCV using the method described above

3.4 Rare and common mutation statistics

I computed simple statistics on the rare and common mutations counts and listed them in **Table 6**.

#mutated positions	#mutations	#rare mutated positions	#rare mutations	#common mutated positions	#common mutations
89	105	74	82	19	23

Table 6. Rare and common mutation statistics across 127 individuals and a 'rare' mutation occurring in <= 3 of them

I have listed some simple statistics for each of the common mutations in Table 7 and 8. 110 files were used in the hypothesis testing. Hence all of the 23 common mutations are not present in the data.

Position	185	263	2617	4264	4529	4580	4769	5513	7507	7507	7509	7509	7526	8295	8295	8297	8297	8303	8348
Mutation	0/1	0/1	0/1	0/1	1/1	1/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/1
Total	1	38	15	19	17	4	18	3	1	15	1	4	3	0	2	2	2	17	9
Average	.03	.97	.38	.49	.44	.10	.46	.08	.03	.38	.03	.10	.08	0	.05	.05	.05	.44	.23

Table 7. Total & average number of occurrences for each of the 20 common mutations for the 39 AD patients

Position	185	263	2617	4264	4529	4580	4769	5513	7507	7507	7509	7509	7526	8295	8295	8297	8297	8303	8348
Mutation	0/1	0/1	0/1	0/1	1/1	1/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/1
Total	3	71	27	42	44	9	22	5	5	26	2	8	5	2	2	2	2	21	20
Average	.04	1.0	.38	.59	.62	.13	.31	.07	.07	.37	.03	.11	.07	.03	.03	.03	.03	.30	.28

Table 8. Total & average number of occurrences for each of the 20 common mutations for the 71 healthy subjects

3.5 Hypothesis testing for common mutations

I have shown the common mutations which were found in the 110 files tested in Table 9 along with their p-values.

Position	185	263	2617	4264	4529	4580	4769	5513	7507	7507	7509	7509	7526	8295	8295	8297	8297	8303	8348
p-value	.66	.18	.96	.3	.06	.71	.11	.90	.32	.85	.94	.87	.90	.29	.54	.54	.54	.14	.57

Table 9. 2 sample t-test results for each of the common mutations. '0' was present in the confidence interval at all the positions

3.6 Genotype and number of mitochondria

Table 10 shows the simple statistics which I computed. **Table 11** shows the results of the 2 sample t-test between estimated number of mitochondria in AD as well as healthy individuals. **Table 12** shows the Spearman's correlation coefficient between the number of mitochondria and the binary vector indicating the presence or absence of a mutation at a position across all cells. The p-value for this correlation is also listed. There are 105 mutations in all so only the relevant coefficients have been shown. A coefficient is considered relevant if its p-value < 0.05.

	total	average
AD	2760333	70777.77
healthy	5048468	71105.19

Table 10. Total and average number of mitochondria across all 110 cells studied in the dataset

p-value	t	Confidence interval
0.95	-0.07	[-9959, 9304]

Table 11. 2 sample t-test for number of mitochondria with AD/NAD

Position	2617	7517	7526	8297	16183	16186	16270	16294	16296	185	189	207	215	250	567	4203
Genotype	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
Common(C)/rare(R)	C	R	C	C	C	R	R	R	R	R	R	R	R	R	R	R
Correlation	0.19	0.19	0.19	0.26	-0.21	0.38	-0.21	0.39	0.24	0.27	0.29	-.21	.19	-.21	.19	.24
p-value	.05	.05	.05	0	.03	0	0.03	0	0.01	0	0	.02	.04	.03	.04	.01

Table 12. Relevant p-values and correlation coefficients for each type of mutation

4. Discussion

4.1 Statistics

Unlike the 0/1 and 1/1 counts (and their averages), the average log allelic ratio appears to give us an indication that mutations are higher in AD patients compared to healthy subjects. Further proof of this is obtained from hypothesis testing. The average age for AD patients is higher which is also in accordance with our hypothesis testing results.

4.2 Hypotheses testing and logistic regression

A p-value under 0.05 tells us that the result is significant. However, for the null hypothesis to be rejected, we require that '0' must not fall in the 95% confidence interval (CI). Thus from **Table 4**, test 1 and 5 is significant. This means that age and log allelic ratios (according to equation 3) have strong correlation with AD. Hence these were chosen as features. This observation also led me to choose equation 3 over 1 or 2 for representing allelic ratios. Test 2,3,4 and 8 include 0 in their CI so I cannot reject the null hypothesis. Test 6 (for gender and AD) has a high p-value but 0 is not in its CI. For the sake of verification, I included it in the logistic regression model but the p-value of 0.7239 showed a less significant though positive (since its coefficient is positive) contribution of gender to AD.

As expected from the hypothesis tests, age and log allele ratio have a significant positive contribution to the presence of AD due their positive estimates and low p-values in the logistic regression model results in **Table 5**.

Test 8 tells us we can ignore the effect of post-mortem interval on the study.

4.3 Rare and common mutations

From **Table 4**, test 2, the total number of mutations in terms of 0/1 and 1/1 genotype counts may or may not have a relation between AD. I propose breaking down these counts into rare and common mutations and testing their relation with AD with separate hypotheses tests. This does not require the use of allelic ratios.

AD patients show a significantly higher rare mutation count compared to healthy individuals (**Table 1**). According to the hypothesis test (Test 9 in **Table 4**) this result is significant. Hence the rare mutation count in AD cells is higher than healthy cells.
Some common mutations have higher average in AD patients while others have higher average in healthy individuals but the t-test sheds further light on this as described below.

4.4 Common mutations hypothesis test

Only the position 4529 (p-value:0.06) seems significant. However since '0' is in the confidence interval at all the positions, the null hypothesis cannot be rejected. In fact, it seems that in some positions (with p-values near 1), the claim to not reject the null hypothesis is high. Thus it seems that Alzheimer's is not related to the presence or absence of a common mutation at a position.

4.5 Number of mitochondria

From the p-values, it seems there is no difference between the number of mitochondria in healthy and diseased cells. Hence AD simply mutates the mitochondrial genome but does not destroy or is not associated with a reduced/increased number of mitochondria.
From the correlation coefficients of **Table 12**, we see that some mutations are associated with an increased number of mitochondria while others are associated with a decreased number of mitochondria. However, the correlation is extremely small in all the cases so though theoretically significant (due to small p-values), the practical significance of these results requires further verification.

4.4 Limitations and future work

The number of subjects is small and hence the hypothesis testing and logistic regression results might improve with more subjects. My current logistic regression model contains log allelic ratios as the only feature obtained from sequencing data. It would be interesting to find more sequencing-features to bolster the performance of logistic regression. However, as the number of features increases, I will require different models or/and more data to achieve high accuracy. It would be interesting to determine if AD leads to a higher mutation count or vice versa. This would require data at various stages of AD. Corroboration with existing studies which identify certain genes as being related to AD could be done. To the best of my knowledge, studies with position and cell level mutation information do not exist which made comparison of this work with existing literature difficult. The mutation counts used in the rare and common mutation experiments could be replaced by our proposed allelic ratios to obtain further insights. Instead of studying the relation between a mutation at a single position and the mitochondria count, it would be interesting to understand the relation between the mt-genotype and the number of mitochondria in the cell.

5. Conclusion

In this work, I used raw RNA-sequencing data from brain cells across individuals of different age and gender to study the relation between mitochondrial mutations and Alzheimer's disease. I tested several hypotheses and established that advanced age, higher cell level mutation (as expressed by allelic ratios) and higher rare mutation count are common in AD cells compared to healthy ones. I built a logistic regression model with ~ 72% accuracy to detect the presence of AD given age, gender and allelic ratio. Such a study helps in studying the pathogenesis of AD and can help in development of treatments and precautionary measures.
AD seems to be associated with an increased load of rare mutations but has little or no relation between the existence or non-existence of a common mutation at a position. Mutations at certain positions influence the number of mitochondria in a cell.

Credit
This project was completed using ideas gleaned from a project under Dr Asa Abeliovich and Dr Herve Rhinn of the Taub Institute for Research on Alzheimer's Disease and the Aging Brain.

References

[1] Flaquer, Antònia, et al. "Mitochondrial GWA Analysis of Lipid Profile Identifies Genetic Variants to Be Associated with HDL Cholesterol and Triglyceride Levels." *PloS one* 10.5 (2015): e0126294.
[2] Wiesner, Rudolf J., J. Caspar Rüegg, and Ingo Morano. "Counting target molecules by exponential polymerase chain reaction: copy number of mitochondrial DNA in rat tissues." *Biochemical and biophysical research communications* 183.2 (1992): 553-559.
[3] Flaquer, Antònia, et al. "Mitochondrial genetic variants identified to be associated with BMI in adults." *PloS one* 9.8 (2014): e105116.
[4] Lambert, Jean-Charles, et al. "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease." *Nature genetics*45.12 (2013): 1452-1458.
[5] Moreira, Paula I., et al. "Mitochondrial dysfunction is a trigger of Alzheimer's disease pathophysiology." *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1802.1 (2010): 2-10.
[6] Montesanto, Alberto, et al. "The Genetic Variability of UCP4 Affects the Individual Susceptibility to Late-Onset Alzheimer's Disease and Modifies the Disease's Risk in APOE-ε4 Carriers." *Journal of Alzheimer's Disease* Preprint (2016): 1-10.
[7] Langmead, Ben, et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." *Genome biology* 10.3 (2009): 1.
[8] Li, Heng. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data." *Bioinformatics* 27.21 (2011): 2987-2993.